AAP Contrats doctoraux en Intelligence artificielle

Cofinancé par l'ANR

LEMUR : LEarning universal MUltimodal Representations of texts and images

1. DESCRIPTION OF THE PHD THESIS PROJECT

1.1 OBJECTIVES OF THE PROJECT BASED ON THE CURRENT STATE OF THE ART

Deep learning has opened a large number of research opportunities in the field of artificial intelligence, leading to breakthroughs in a lot of related tasks, in particular natural language processing and computer vision.

Beyond convolutional and recurrent models, multilayer-multihead attention mechanisms are establishing as a major component of deep learning systems, for their capability to model complex causal relations between non-adjacent inputs. These models, based on the transformer architecture [Vaswani 2017], developed in the natural language processing community, are typically hard to train on small fully supervised datasets. But they can be pre-trained on self-supervised tasks, such as guessing hidden words for text, or regenerating redacted regions of images, on very large unlabled datasets, and subsequently fine-tuned on supervised tasks with great success. The example of BERT [Delvin 2018] has been a game changer in the NLP community, improving state of the art performance over a number of difficult NLP tasks, and generating a copious number of variation studies exploiting the same ideas [Yang 2019, San 2020, Liu 2019...]. In the artificial vision community, similar ideas first developped through the automatic captioning task which is similar to machine translation, but in combination with convolutional layers in order to reduce the computational cost associated with attention [You 2016], and later extended to other artificial vision tasks [Pamar 2018].

One central question of this proposal is whether multilayer-multihead attention mechanism can produce universal representations of multimodal linguistic phenomena, in the same way it induced breakthroughs in the natural language community. In a preliminary study, the TALEP group at LIS has developped corpora and systems for syntatctic parsing of text grounded in an image described by that text [Delecraz 2019]. This study has shown that joint modeling of text and images can help NLP systems generalize in the context of evidence coming from non-textual modalities. Other studies have started to explore the application of multilayer-multihead attention mechanisms to multimodal content. In particular, most relevant to this project, the UNITER framework aims at training a single attention-based encoder on top of representations extracted

from texts and images [Chen 2019]. As for its unimodal counterpart, the model is pretrained on self-supervised tasks including cross-modal tasks such as predicting a word from the image or predicting an image region from the text. Such models, when fined-tuned on text-image tasks, such as visual question answering or cross-modal natural language inference, leads to performance improvement over state of the art systems.

Given this context, it is now a good time to explore a number of challenges related to the emergence of universal multimodal representations:

- How to address the low quantity of aligned cross-modal data available, even for training self-supervised tasks?
- What is so special with attention mechanisms that they require fine-tuning of whole models and therefore do not allow for universal representations?
- What type of self-supervision can be effective in cross-model context when little or no alignement data is available.

These challenges will be explored through a careful analysis of attention mechanisms, the introduction of novel adversarial self-supervision schemes [Goodfellow 2014], and through the development of original self-supervision tasks based on existing unimodal semantic representations [e.g. Caron 2018, Gidaris 2018].

[Vaswani 2017] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.
[Delvin 2018] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
[Yang 2019] Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." Advances in neural information processing systems. 2019.
[San 2020] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).
[Liu 2019] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).
[Pamar 2018] Parmar, Niki, et al. "Image transformer." arXiv preprint arXiv:1802.05751 (2018).
[You 2016] You, Quanzeng, et al. "Image captioning with semantic attention." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[Chen 2019] Chen, Yen-Chun, et al. "Uniter: Learning universal image-text representations." arXiv preprint arXiv:1909.11740 (2019).
[Delecraz 2019] Delecraz, Sebastien, et al. "Visual Disambiguation of Prepositional Phrase Attachments: Multimodal Machine Learning for Syntactic Analysis Correction." International Work-Conference on Artificial Neural Networks. Springer, Cham, 2019.
[Caron 2018] Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[Gidaris 2018] Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." arXiv preprint arXiv:1803.07728 (2018).
[Goodfellow 2014] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

1.2 METHODOLOGY

During the project, the candidate will build and evaluate various multimodal representation learning strategies. These systems will be developped using standard tools adequate for deep learning research (typically Pytorch, Tensorflow).

The candidate will use the various datasets of curated at LIS or in the community for joint image and text modelling, including the COCO dataset (http://cocodataset.org), the visual genome dataset (http://visualgenome.org), and various large scale unimodal datasets such as WebAsACorpus (WaC) or Imagenet.

The systems will be evaluated on a range of multimodal text-image tasks:

- Image captioning: COCO, Flickr
- Finegrained relation generation: Visual genome
- Grounding: NLVR
- Visual question answering (VQA)
- Visual Commonsense Reasoning (VCR)
- Image-text and text-image retrieval
- Visual entailment (SNLI-VE)
- Referring Expression Comprehension (RefCOCO)
- Syntactic : prepositionnal phrase attachment (pp-Flickr)

1.3 WORK PLAN

Month 0-6: survey of relevant material in the field => publication

Month 6-12: development of baseline multimodal representation system

Month 12-18: development of self-supervision tasks => publication

Month 18-24: development of adversarial approaches => publication

Month 24-30: development of universal representations => publication

Month 30-36: PhD writing

1.4 SUPERVISOR AND RESEARCH GROUP DESCRIPTION

The thesis project will be advised by members of the TALEP (natural language processing; https://www.lis-lab.fr/en/talep-2/) and QARMA (machine learning; https://www.lis-lab.fr/en/qarma/) teams of LIS, under an ongoing effort on multimodal language processing, involving B. Favre, L. Becerra, A. Nasr, T. Artieres, S. Ayache, R. Sicre.

The supervisor is B. Favre (https://pageperso.lis-lab.fr/benoit.favre/), MCF HDR in the TALEP group, co-head of the multimodal language axis of the CNRS GDR TAL (https://gdr-

tal.ls2n.fr/), and head of data science at LIS. The thesis will be co-supervised by a member of QARMA.

The candidate will be able to use the computational infrastructure of LIS, the AMU mesocentre, and the Jean-Zay cluster from CNRS for accelerating large scale deep learning computation.

2. RECENT PUBLICATIONS

Sebastien Delecraz, Leonor Becerra-Bonache, Alexis Nasr, Frederic Bechet, Benoit Favre, "Visual Disambiguation of Prepositional Phrase Attachments: Multimodal Machine Learning for Syntactic Analysis Correction", International Work-Conference on Artificial Neural Networks (IWANN), 2019

Gabriel Marzinotto, Géraldine Damnati, Frédéric Béchet, Benoit Favre, "Robust Semantic Parsing with Adversarial Learning for Domain Generalization", Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers), 2019

3. EXPECTED PROFILE OF THE CANDIDATE

The candidate should:

- be passionate about the various topics of artificial intelligence

- have a strong academic record in computer science and engineering or mathematics

- have specialized at the MSc level in machine learning with applications to natural language processing or computer vision

- be rigorous, tenacious, and have strong scientific skills

- enjoy writing about science

- have good knowledge of programming in python and at least one lower-level language (an experience in developing deep learning systems with current technology is a plus)

4. SUPERVISORS' PROFILE

Benoit Favre

- Web: https://pageperso.lis-lab.fr/benoit.favre

Positions

- PhD on "Automatic Summarization of Speech", Computer Science, Avignon University (UAPV), France (2007)
- Postdoc: International Computer Science Institute (NLP & machine learning), Berkeley, USA (2007-2009); LIUM Université du Maine (2009-2020)
- Associate Professor (Maître de Conférences), Aix-Marseille University, France (2010-)
- Research Scholar, Queensland University of Technology, Australia (2018-2019)

Themes

- Natural language processing and machine learning, multimodal and spoken language understanding

Community service

- Co-head of data science at Laboratoire d'Informatique et Systèmes (LIS)
- Co-head of MSc on artificial intelligence and machine learning at AMU
- Co-head of multimodality axis at CNRS GDR TAL
- Associate editor of Revue TAL
- Board member of ATALA (French NLP society)
- Member of ILCB, Archimede

Projects

- 1 ERC (participant, PI C. Henriot), one EU FP7 project, and 3 DARPA projects, 8 French ANR projects, 1 industry (Orange Labs), 1 Amidex, 1 CEFIPRA (International, India), 1 ARC (Australia)

Students

- 4 PhD graduated
- 4 PhD running (started 2016, 2017, 2017, 2019)

Publications

- More than 110 publications, among which 4 book chapters, 9 peer-reviewed international journals, 64 peer-reviewed international conferences, h-index of 22

VISA DU RESPONSABLE DE L'INSTITUT ET DU DIRECTEUR DE LABORATOIRE CONCERNÉS

Visa du responsable de l'institut, **NOM Prénom**	**Visa du directeur du laboratoire,** **NOM Prénom**
GODARD Emmanuel	
Fait à Marseille, le 15/05/2020	**Fait à Marseille, le**
Signature 	**Signature**